# CCL Classification Rules and Models

Michael Messner[1], Zeno Bain[1], Thomas Carpenter[1], Yvette Selby-Mohamadu[1], Joyce Donohue[1], Wynne Miller[1], Frank Letkiewicz[2], JoAnne Shatkin[2], and George Hallberg[2]

[1] USEPA, Office of Water, Washington, DC, USA.

[2] The Cadmus Group, Watertown, MA, USA.

## BACKGROUND

EPA plans to use classification rules and models to help identify contaminants for its Contaminant Candidate List (CCL). A training data set (TDS) is used to calibrate or "train" the rules/models.

The TDS (described in previous poster) is a set of contaminants, each scored for its health effects (Potency and Severity), and occurrence attributes (Prevalence and Magnitude) and assigned to one of four categories: Not List (1), Not List? (2), List? (3), and List (4). Also available are average classifications, such as 3.5 when half the experts assigned the contaminant to List and half assigned it to List?

Classification algorithms/models considered are:

- Artificial Neural Networks (ANN)
- Classification and Regression Tree (CART®)
- Multivariate Adaptive Regression Splines (MARS®)
- Quick, Unbiased, Efficient Statistical Tree (QUEST®)
- Simple linear model

## PRINCIPLES and OBJECTIVES

A good rule/model/algorithm would:

- Correctly predict most of the training data set classifications. Errors should be avoided.
- Avoid the most serious errors. Produce small total error loss (see loss table below).

| Rule-Based Decision | Original (Team) Decision | | | |
|---|---|---|---|---|
| | Not List | Not List? | List? | List |
| Not List | -- | 2 | 5 | 10 |
| Not List? | 1 | -- | 2 | 5 |
| List? | 2 | 1 | -- | 2 |
| List | 3 | 2 | 1 | -- |

Ideally, the algorithm used would also:

- be clear (rather than hidden), based on non-proprietary software
- work with missing data (missing score for an attribute)
- provide some continuous measure of strength, so contaminants can be sorted

## PERFORMANCE ISSUES

**CART**:

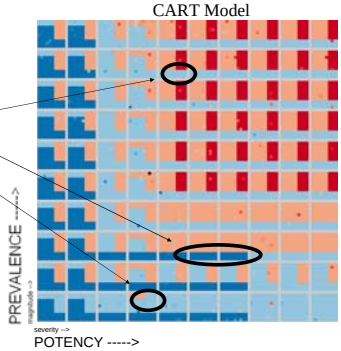
•Choppy (vertical and horizontal discriminants)

•NL touches L? and L touches NL?

•A few spots where increasing on attribute decreases the classification.

**MARS** (not shown): Many discriminants are horizontal or vertical. Discrepancies are less numerous and more difficult to spot.

CART Model




**ANN**:

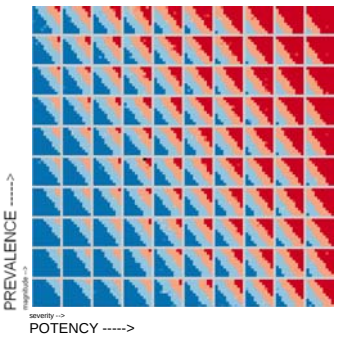
•Diagonal discriminants make sense.

•Consistent ordering of classifications (NL? always separates NL from L? and L? always separates NL? from L)

•Classification always increases with increasing attribute scores.

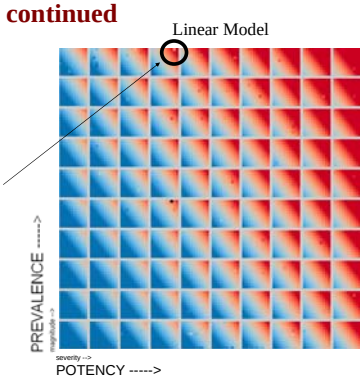
**Linear and QUEST** (not shown): Look similar to ANN. No inconsistencies. QUEST has the most red, to minimize error loss. Linear and ANN had other objectives (maximize likelihood and minimize error count).

ANN Model




## PERFORMANCE ISSUES, continued

Linear Model




Linear model predicts average team decision. Using smooth color gradient (rather then 4 colors) reveals quality of fit to training data set. Notice that relatively few of the 202 training set contaminants stand out.

Largest "error" was a sodium compound (4, 9, 10, 10), which the team declared L? All five models predicted L for this point.

Error Loss

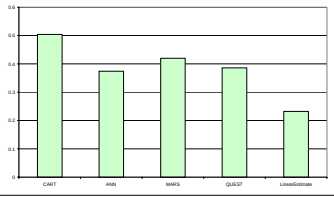
QUEST, trained to minimize error loss, did the best job of avoiding error loss. The linear model did not do well because it attempted to predict team average. When team average was, say 3.5, the "rounded" decision was List, but the linear model often predicted L? and accrued some error loss.


Mean Squared Error

The linear model, designed to minimize this kind of error, did it best.

CART and MARS performed poorly with respect to both error loss and mean squared error.

ANN performance appears satisfactory.



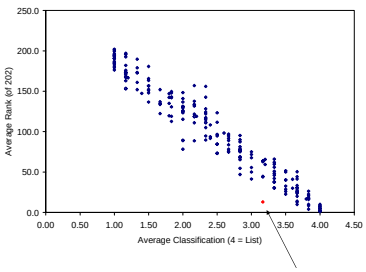

## COMBINING INFORMATION FROM MULTIPLE RULES/MODELS

Rank Contaminants by "Strength" for Listing:

- For ANN, sort by probability of membership in List set
- For linear model, sort by estimated team average classification
- For QUEST, sort by distance from nearest neighbor in next higher (or lower) class.

Combine Info from three models:

1. Sort and find contaminant rank by the three methods (ANN, QUEST, linear).
2. Average the ranks.
3. Sort by average rank. The figure below shows that this tracks well with team average classification.




(sodium salt with very high occurrence)

## REFERENCES

Classifying Drinking Water Contaminants for Regulatory Consideration, National Research Council, Committee on Drinking Water Contaminants, NRC Press, 2001.

National Drinking Water Advisory Council Report on the CCL Classification Process, 2004.

Graph colors based on www.ColorBrewer.org, by Cynthia A. Brewer, Penn State Univ.